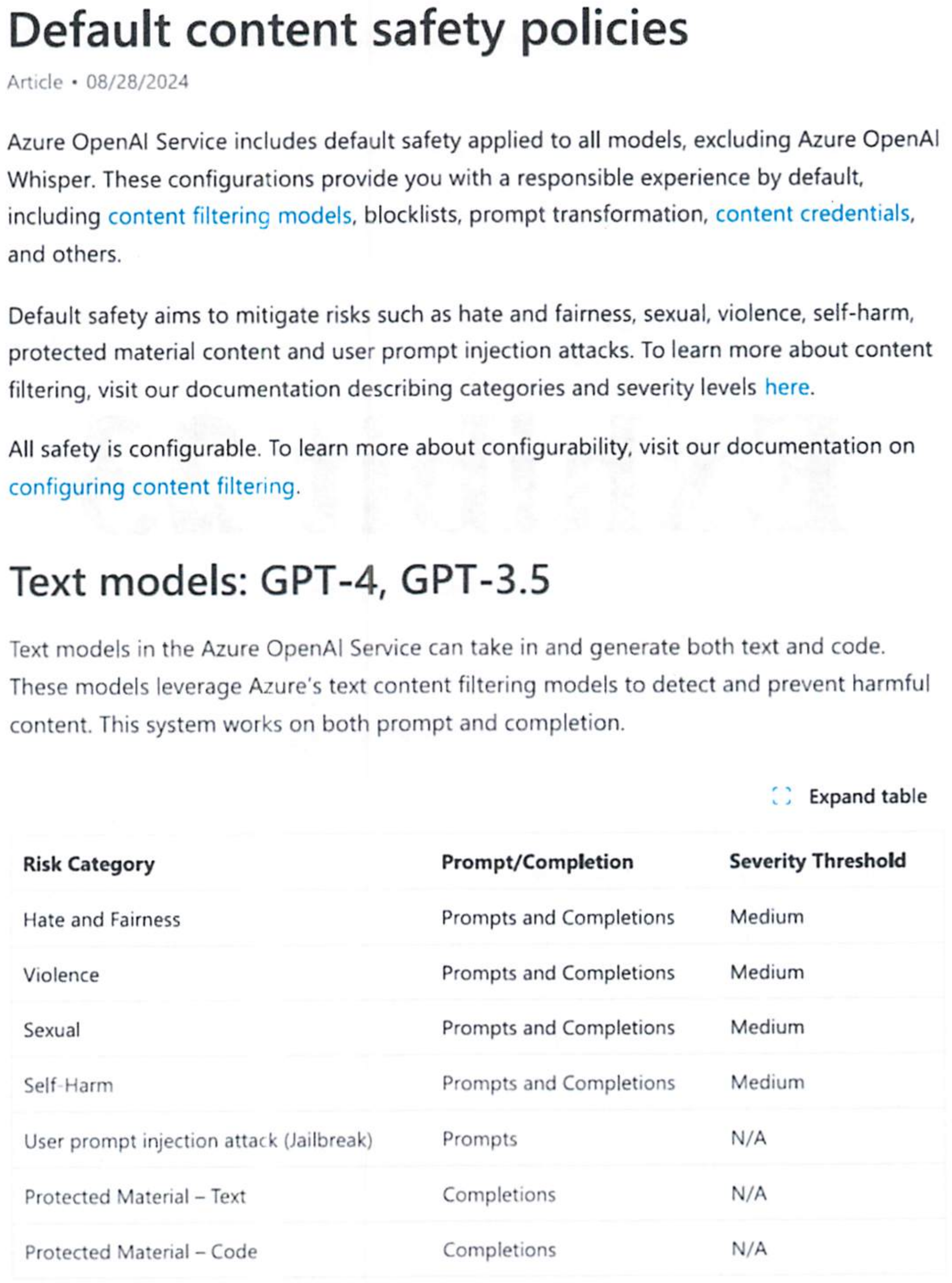
Exhibit 23

https://learn.microsoft.com/en-us/azure/ai-services/openai/concepts/default-safety-policies



Default content safety policies

Article • 08/28/2024

Azure OpenAI Service includes default safety applied to all models, excluding Azure OpenAI Whisper. These configurations provide you with a responsible experience by default, including content filtering models, blocklists, prompt transformation, content credentials, and others.

Default safety aims to mitigate risks such as hate and fairness, sexual, violence, self-harm, protected material content and user prompt injection attacks. To learn more about content filtering, visit our documentation describing categories and severity levels here.

All safety is configurable. To learn more about configurability, visit our documentation on configuring content filtering.

Text models: GPT-4, GPT-3.5

Text models in the Azure OpenAI Service can take in and generate both text and code. These models leverage Azure's text content filtering models to detect and prevent harmful content. This system works on both prompt and completion.

Risk Category	Prompt/Completion	Severity Threshold
Hate and Fairness	Prompts and Completions	Medium
Violence	Prompts and Completions	Medium
Sexual	Prompts and Completions	Medium
Self-Harm	Prompts and Completions	Medium
User prompt injection attack (Jailbreak)	Prompts	N/A
Protected Material – Text	Completions	N/A
Protected Material – Code	Completions	N/A

Vision models: GPT-4o, GPT-4 Turbo, DALL-E 3, DALL-E 2

GPT-4o and GPT-4 Turbo

Expand table

Risk Category	Prompt/Completion	Severity Threshold
Hate and Fairness	Prompts and Completions	Medium
Violence	Prompts and Completions	Medium
Sexual	Prompts and Completions	Medium
Self-Harm	Prompts and Completions	Medium
Identification of Individuals and Inference of Sensitive Attributes	Prompts	N/A
User prompt injection attack (Jailbreak)	Prompts	N/A

DALL-E 3 and DALL-E 2

Expand table

Risk Category	Prompt/Completion	Severity Threshold
Hate and Fairness	Prompts and Completions	Low
Violence	Prompts and Completions	Low
Sexual	Prompts and Completions	Low
Self-Harm	Prompts and Completions	Low
Content Credentials	Completions	N/A

Deceptive Generation of Political Candidates	Prompts	N/A
Depictions of Public Figures	Prompts	N/A
User prompt injection attack (Jailbreak)	Prompts	N/A
Protected Material – Art and Studio Characters	Prompts	N/A
Profanity	Prompts	N/A

In addition to the above safety configurations, Azure OpenAI DALL-E also comes with prompt transformation by default. This transformation occurs on all prompts to enhance the safety of your original prompt, specifically in the risk categories of diversity, deceptive generation of political candidates, depictions of public figures, protected material, and others.

Feedback

Was this page helpful? Yes No

Provide product feedback | Get help at Microsoft Q&A